The spatial and social clustering and diffusion of vaccine refusal: Analysis Notes, Planning, and Summary Document

Paul Delamater, Varun Goel, Kelsey Sumner, and Natalie Smith

Created 06/01/2018

Table of Contents

Last Updated 01/14/2019

# Aim #1

**Evaluate the extent to which clusters of vaccine refusers are defined by geographic proximity, socio-demographic similarity, commuting patterns, or a combination of the them.**

**Approach:** Implement global and local autocorrelation methods per [Delamater et al. 2018, Examining the spatiotemporal evolution of vaccine refusal: nonmedical exemptions from vaccination in California, 2000-2013](https://doi.org/10.1186/s12889-018-5368-y) using various neighborhood definitions. Compare the results of various neighbor definitions to help understand which of them best explains “clustering” of similar behaviors. Test each year’s NME data over 2000-2013.

**Data aggregation:** In previous research, used School, Block Group (BG), Tract, and School District (SD). Results of previous research focused on Tract-level results. Potentially implement each level of aggregation, but likely start with BG and Tract-level data for comparative purposes, and then move onto School and SD.

## Geographic Proximity

Already completed in Delamater et al 2018.Used K=5 nearest neighbors for all tests. NME spatiotemporal dataset already created (for all years and all levels of data aggregation). The R code to conduct spatial autocorrelation tests and to consolidate results is already written.

* Code at: vaccine\_exemptions/code/ref

## Socio-demographic similarity **[Kelsey]**

* ~~Determine social, economic, and demographic attributes to be used to evaluate similarity from previous literature on refusal/exemptions~~
* ~~Gather/clean social, economic, and demographic attribute data~~
  + Data at: OneDrive/CA\_NME\_Diffusion/data/soc\_econ\_dem/yearly
    - Race/Ethnicity (8 variables, proportions), Age (18 variables, proportions), Education (5 variables, proportions), Income (2 variables, dollars)
* ~~Determine metric to evaluate similarity among places based on multiple variables (Probably Euclidean distance, but maybe RMSE)~~
  + Notes:
    1. *Weigh similarity of Race/Eth, Age, Education, and Income equally*
    2. *Must consider scale of input variables measured in different units (e.g., proportions versus dollars) and how overall concepts are scaled (9 variables vs 2 variables), as well as the metric output (e.g., what is the similarity metric result for perfectly similar and for perfectly dissimilar regions?)*
    3. *For Age and Education, calculate Cumulative Sum of proportions first (from lowest group to highest group), as this will allow us to evaluate similarities/differences “within” the variables*
    4. *Use Euclidean distance to calculate similarity scores*
  + See example/explanation at: docs/SimilarityExample.Rmd
  + Code for creating cumulative sum of proportions: code/analysis/Kelsey/social\_analyses/01\_create\_cum\_prop\_ordinal\_vars\_KS.R
  + Code for calculating similarity scores: code/analysis/Kelsey/social\_analyses/02\_calculate\_similiarty\_scores\_soc\_ecom\_dem\_KS.R
  + Sample of histograms of similarity score distributions can be found at: docs/Similarity Score Distributions.pdf
  + Data showing the matrices with the cumulative proportions for ordinal variables: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> soc\_econ\_dem -> cumprop\_yearly
  + Data showing the matrices with the similarity scores: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> soc\_econ\_dem -> SI\_yearly
* ~~Determine which year of data will be used – have all temporal data~~
  + Notes:
    1. *Begin with a single year (2016), and using matching years (PBE to Community attributes).*
    2. *We had sociodemographic data for 2000 to 2016 but PBE For 2000 to 2015, so matched output for years 2000 to 2015.*
* ~~Construct neighborhood weight matrix in R to express similarity among regions~~
  + Notes:
    1. *Constrain to only observations with Enrollment data in each year*
    2. *Use K= 5 nearest neighbors to match spatial test*
    3. *See Varun’s code to format for spdep functions*
  + Code for calculating K=5 nearest neighbors: code/analysis/Kelsey/social\_analyses/03\_calculate\_KNN5\_soc\_econ\_dem\_KS.R
  + Data showing the K=5 nearest neighbors results: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> soc\_econ\_dem -> KNN5\_yearly
* ~~Run global and local spatial autocorrelation tests over 2000-2013 using PBE data and gather/organize/consolidate output~~
  + Notes:
    1. *For final output, constrained data set to years 2002 to 2015, which have LODES, PBE, and sociodemographic data.*
  + Code for calculating local spatial autocorrelation: code/analysis/Kelsey/social\_analyses/04\_calculate\_LISA\_soc\_econ\_dem\_KS.R code/analysis/Kelsey/social\_analyses/05\_consolidate\_LISA\_results\_soc\_econ\_dem\_KS.R
  + Code for summarizing local spatial autocorrelation results in tables: code/analysis/Kelsey/social\_analyses/07\_analyze\_LISA\_results\_soc\_econ\_dem\_KS.R
  + Data showing LISA results for K=5 nearest neighbors and summary tables: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> soc\_econ\_dem -> LISA\_results
  + Code for calculating global spatial autocorrelation: code/analysis/Kelsey/social\_analyses/06\_calculate\_global\_morans\_I\_soc\_econ\_dem\_KS.R
  + Data showing moran’s I results for K=5 nearest neighbors: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> soc\_econ\_dem -> global\_morans\_I\_results
* Assess overlap in K=5 nearest neighbors between social and spatial data sets by creating histograms
  + Code for calculating K=5 nearest neighbors for spatial data set: code/analysis/Kelsey/spatial\_analyses/misc\_creating\_KNN5\_for\_spatial\_data\_KS.R
  + Data showing spatial data K=5 nearest neighors results at (not used for the combined results - just for creating histograms of social overlap: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> spatial\_data -> KNN5\_yearly
  + Code for assessing overlap between social and spatial data sets: code/analysis/Kelsey/social\_analyses/misc\_create\_histograms\_of\_KNN5\_overlap\_KS.R

## Commuting Data **[Varun]**

* ~~Use LODES data to determine links among regions (Block Group to Block Group and Tract to Tract) based on proportion of people travelling among regions for work~~
* ~~Construct neighborhood weight matrix in R to express similarity among regions~~
  + Notes:
    1. *~~Constrain to only observations with Enrollment data in each year~~*
    2. *~~Use K= 5 nearest neighbors to match spatial test~~*
  + Code at: vaccine\_exemptions/code/analysis/03\_calculate-lisa-lodes\_VG.R
* ~~Run global and local spatial autocorrelation tests over 2000-2013 using PBE data and gather/organize/consolidate output~~ \*~~Create K=5 nn for affiliation networks using one-mode bipartite projection of lodes data~~
* Notes:
  1. *~~Include ties and assign weights based on ties~~*
* Data at: (tracts) data/results/nblist\_bipartite\_tracts\_2002\_2015.RData
* Data at: (block groups) data/results/nblist\_bipartite\_bg\_2002\_2015.RData
* ~~Run global and local spatial autocorrelation tests over 2002-2015 using PBE data and gather/organize/consolidate output~~
* Data at: (tracts) data/results/CA\_tct\_2010\_wLisaBipartiteResults.csv
* Data at: (block groups) data/results/CA\_bg\_2010\_wLisaBipartiteResults.csv
* Shapefile at: (both tracts and block groups) data/results/CA\_LISA-bipartite-Allyears\_shp\_VG.zip

## Combination of Geographic, Social, Commuting **[Kelsey]**

* ~~Construct neighborhood weight matrix in R to express~~ **~~combined~~** ~~similarity among regions~~
  + Notes:
    1. *Not combining all three types of data now - only the geographic and social data for similarity*
    2. *Constrain to only observations with Enrollment data in each year*
    3. *Calculate spatial weights using Euclidean distance in space from the shapefiles for BG and TCT*
    4. *Match social similarity and spatial weights matrices*
    5. *Consider how each element can be weighted in the combination, e.g., geography (50%) and socio-demographic (50%)*
    6. *Combining weights for geographic and social data over years 2002 to 2015 for consistency with years available in all data sets*
    7. *Calculate K=5 Nearest Neighbors for the combined social and spatial weights*
  + Code for calculating spatial weights using Euclidean distance in space: code/analysis/Kelsey/spatial\_analyses/01\_create\_spatial\_weights\_KS.R
  + Data showing Euclidean distance in space for spatial data set: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> spatial\_data -> Euclidean\_dist
  + Code for combining spatial and social weights at: code/analysis/Kelsey/combo\_social\_spatial\_analyses/02\_combine\_weights\_social\_and\_spatial\_KS.R
  + Data showing combine social and spatial weights: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> combined\_social\_spatial -> weights\_matrices\_yearly
  + Code for combining spatial and social weights at: code/analysis/Kelsey/combo\_social\_spatial\_analyses/03\_calculate\_KNN5\_combo\_data\_KS.R
  + Data showing combine social and spatial weights: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> combined\_social\_spatial -> KNN5\_yearly
* ~~Run global and local spatial autocorrelation tests over 2002-2015 using PBE data and gather/organize/consolidate output~~
  + Code for calculating local spatial autocorrelation: code/analysis/Kelsey/combo\_social\_spatial\_analyses/04\_calculate\_LISA\_combined\_social\_spatial\_data\_KS.R code/analysis/Kelsey/combo\_social\_spatial\_analyses/05\_consolidate\_LISA\_social\_spatial\_combo\_KS.R
  + Data showing local spatial autocorrelation results: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> combined\_social\_spatial -> LISA\_results -> LISA\_shp\_all
  + Code for summarizing local autocorrelation results: code/analysis/Kelsey/combo\_social\_spatial\_analyses/07\_analyze\_LISA\_results\_social\_spatial\_combined\_KS.R
  + Data summarizing spatial autocorrelation results: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> combined\_social\_spatial -> LISA\_results -> LISA\_tables
  + Code for calculating global autocorrelation: code/analysis/Kelsey/combo\_social\_spatial\_analyses/06\_calculate\_global\_morans\_I\_social\_and\_spatial\_KS.R
  + Data showing global autocorrelation results: OneDrive -> CA\_NME\_Diffusion -> kms\_data -> combined\_social\_spatial -> global\_morans\_I\_results

## Comparison of Spatial Autocorrelation results

* Comparison of the neighborhood definitions themselves, e.g., how similar is the “set” of neighborhoods among the different approaches? Initial idea is to calculate a proportion of “matches” based on the original, spatial neighbors and to create frequency tables of matches (e.g., X obs matched all 5, Y obs matched 4, Z obs matched 3, …)
* Global autocorrelation: Should be simple, evaluate magnitude of Moran’s I over time. Since the input data are the same, the output of the different neighborhood results are directly comparable.
* Local autocorrelation: Will have to think about local autocorrelation results… e.g., number of HH or LL observations? Overlapping geography?… even though neighborhoods are different, the output is at an observation level.

# Aim #2

**Develop a preliminary set of models to disentangle the effects of space, interaction patterns, and socio-demographic similarity on year-to-year changes in vaccine refusal.**

**Approach:** Explore and examine methods and techniques to measure “diffusion” of exemptions over 2000-2013. May require developing our own technique.

## Systems Dynamics Modeling **[Natalie]**

* ~~Examine general applicability of SD model (with spatiotemporal observations) to evaluate diffusion~~
  + The approach would generally be: have one stock of ‘unvaccinated/exempted’ and one stock of ‘vaccinated/non-exempted’ for each census tract. Those tracts would then be connected with spatial, social, and commuting data. After the model structure was built, we could calibrate the spatial/social/commuting connections to get some sort of ‘effect size’ for how connected tracts influence each other.
  + Discussed with group (and Leah Frerichs) and although the actual idea is feasible, it would require:
  1. Version of Vensim Natalie doesn’t have access to that could pull in data from excel and perform ‘subscript’ operations for each tract
  2. Intense compartmental modeling coding in R to create the structure, and then further work to calibrate the model
  3. Is also likely not the best modeling option even with the appropriate software, because other types of simulation models such as agent-based models allow for socially and spatially explicit connections and are designed to be for this type of process

## Network Diffusion **[Natalie]**

* Examine general applicability of using the netdiffuseR package to evaluate diffusion
  + Notes:
    1. Discussion of a “threshold” for adoption (5/31/18), we will likely start using 5%
    2. There is a lot of change in adoption - e.g., someone could adopt one year, then deadopt the next year
       - Currently there is no good way to handle this in netdiffuseR (their developer suggested EpiModel)
       - For now, will use the time of first adoption
    3. Per the Valente et al 2015 article, (Valente, T. W., et al. (2015). “Diffusion of innovations theory applied to global tobacco control treaty ratification.” Social Science & Medicine 145: 89-97.), it would be possible to hypothesize perfect diffusion and then compare with what we actually see
       - Will read the paper more closely and evaluate next steps analytically - begin now with complete case analysis on tracts who have pber data for 2002-2015 and use the lodes network

## Agent-Based Modeling **[Natalie]**

* Examine general applicability of ABM to evaluate diffusion over time
  + Notes:
  1. Some type of ABM would allow us to place the census tracts in geographic space, explicitly define who they are connected to in various ways (spatial, social, and commuting), and input their vaccine data over time. The model could then be calibrated (like discussed above for the SD modeling) to get a potential effect size for the diffusion parameter
  2. A SIENA model is an ABM that has the added layer of statistical inference. Natalie is attending a conference at the beginning of June and will consider the potential use of a SIENA model (which would allow for inference vs. just calibration).
  3. Siena model has some assumptions / requirements that make it less useful for our work
     + could not handle 5551 census tracts computationally – would have to analyze in chunks at the county or regional level
     + assumes that every node in the network could theoretically be connected to other nodes – not reasonable
     + is built on the framework that the network structure and behavior variables coevolve over time
     + we have network change – but it’s network change that’s the result of missing data, not because the network is changing in response to or because of other variables (like vaccination)
     + per 6/22 discussion between PD and NS, NS will focus on diffusion modeling re: Valente et al 2015 article

## Perfect Diffusion Model **[Natalie]**

* Construct a dataset that has perfect diffusion over time. The idea here is that, eventually, this might be able to be used to create/develop a inferential test on diffusion, given observed data.
  + Notes:
  1. Can use a “toy” set of observations, or actual CA data
  2. Make code flexible, so can eventually be parameterized from observed data
  3. Requred parameters: Seed locations (e.g., which observations have high NME), diffusion paramter (the NME% that is spread/transfered from seeds to neighbors)
  4. Conceptual approach: Identify seed locations at time = t, identify neighbors of seed locations at time t+1, add NME% to neighbors of seed locations at time t+1, repeat
* Ideas for later: Use observed NME attributes to parameterize the model (e.g., auto-extract from observed, then plug into perfect diffusion model); Run basic spatial autocorrelation tests (Moran, LISA, Getis) on perfect diffusion output (e.g., to evaluate what these look like in a perfect diffusion setting); test how observed data deviates from “perfect diffusion”

# Defining Herds

**Rationally define “herds” for better understanding herd immunity and disease outbreak risk via the use of vaccination coverage, R0, the herd immunity threshold, and spatial clustering/arrangement.**

Since we’re working on multiple ways to define neighborhoods and links among places, I want to generally think about how we can integrate them to create regions that are strongly connected (via space, socio-econ-demo, and travel) that would constitute a herd. There’s a large literature in geography on regionalization (and grouping) for us to draw from, but no one has really linked this with the idea of herd immunity and outbreak risk.

* See poster for initial reference: OneDrive/CA\_NME\_Diffusion/docs/ref/Delamater-ACVR-160418-sm.pdf
* This is where, “analysis of the network itself” might come in really handy, e.g., examine the most important links among places to help determine where to “separate” nodes into groups of nodes.